Les applications industrielles de la bio-informatique

Jean-Philippe Vert

▶ **To cite this version:**

Jean-Philippe Vert. Les applications industrielles de la bio-informatique. Réalités industrielles. Annales des mines, 2013, Février 2013, pp.17-23. hal-00796732

HAL Id: hal-00796732

https://hal-mines-paristech.archives-ouvertes.fr/hal-00796732

Submitted on 4 Mar 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Les applications industrielles de la bio-informatique[1]

Jean-Philippe Vert

Directeur du Centre de Bio-informatique à MINES ParisTech

Jean-Philippe.Vert@mines.org

Résumé

A l'heure où les technologies à haut débit en génomique et en protéomique envahissent les laboratoires de biologie, les sciences de la vie font face à un déluge de données extraordinairement volumineuses et complexes. Manipuler ces données et en extraire un sens biologique requiert de nouvelles approches basées sur la modélisation et l'informatique. La bio-informatique est tout à la fois une science, à l'interface entre l'informatique et la biologie, et une industrie vitale pour stocker, diffuser, analyser et interpréter les données biologiques en vue de leur exploitation dans l'industrie de la santé, l'agro-alimentaire ou l'énergie. Cet article propose un rapide tour d'horizon du domaine, de ses acteurs et de ses défis.

Biographie

Diplômé de l'Ecole Polytechnique (X92) et du corps des Mines, Jean-Philippe Vert a obtenu une thèse de mathématique à l'Ecole normale supérieure de Paris avant de travailler à l'Université de Kyoto (Japon) puis à l'Ecole des Mines de Paris, où il a créé le Centre de Bio-informatique qu'il dirige toujours. Il est également directeur adjoint d'une unité mixte de recherche entre Mines ParisTech, l'Institut Curie et l'INSERM sur la bio-informatique du cancer. Il est expert en apprentissage statistique et en bio-informatique, et s'intéresse en particulier à leurs applications dans la recherche contre le cancer. Auteur de plus de 80 publications scientifiques, il est lauréat de la médaille de bronze du CNRS et d'une bourse du Conseil Européen de la Recherche.

[1] Version 27/11/2012. A paraître dans *Annales des Mines - Réalités Industrielles*, p.17-23, février 2013.

Les progrès technologiques fulgurants des 20 dernières années ont eu un impact radical sur les sciences du vivant. Les techniques de puces à ADN, de protéomique, d'imagerie ou maintenant de séquençage se sont invitées au cœur des laboratoires, offrant aux scientifiques de nouveaux outils pour scruter et quantifier le vivant dans ses moindres détails. Il est maintenant possible de mesurer au niveau moléculaire l'ensemble des modifications génétiques portées par les cellules cancéreuses d'un patient, pour tenter d'y répondre par un traitement complètement personnalisé [1]; de quantifier les millions de micro-organismes évoluant dans un milieu naturel particulier, pour identifier de nouvelles manières de transformer la matière [2]; ou d'observer les changements morphologiques induits par la suppression de chacun des milliers de gènes d'un organisme donné par vidéo-microscopie, pour identifier de nouvelles cibles thérapeutiques [3]. Les possibilités offertes par l'utilisation de ces nouvelles technologies pour étudier et comprendre le vivant ne semblent avoir de limite que notre imagination. De par leur caractère systématique et quantitatif, elles ouvrent notamment la voie à de nouvelles approches quantitatives pour modéliser le vivant. Elles posent cependant de nouveaux défis qui dépassent largement le cadre traditionnel des sciences du vivant, de par les vastes quantités de données complexes qu'elles génèrent. Par exemple, le projet américain TCGA (« The Cancer Genome Atlas ») qui vise à cataloguer les variations génomiques de 10,000 cancers, génère 10 téraoctets de données chaque mois, et devrait en produire un total de 10 pétaoctets[2]. Pour transmettre, stocker, analyser et interpréter ces données, afin d'aboutir à un résultat biologique, la biologie s'appuie naturellement de plus en plus sur les mathématiques et les technologies de l'information et de la communication. Une discipline nouvelle est d'ailleurs née pour répondre à ces défis depuis une quinzaine d'année : la bio-informatique, qui s'épanouit à la fois comme un domaine scientifique à part entière, et comme un ensemble de technologies devenues indispensables pour la recherche académique et industrielle. Dans les quelques pages qui suivent, je vais tenter de mieux définir la discipline et d'illustrer ses principaux domaines d'application, en me concentrant essentiellement sur les applications industrielles. Je brosserai ensuite un rapide portrait du marché et des acteurs de ce secteur, avant de conclure par quelques pistes de réflexion sur les défis qui s'offrent à nous et qui représentent autant d'opportunités scientifiques et industrielles.

Qu'est-ce que la bio-informatique ?

La bio-informatique est une discipline à l'interface de la biologie et de l'informatique, qui recouvre l'ensemble des technologies et des méthodes permettant de collecter, de stocker, d'analyser et d'interpréter les données biologiques. Elle désigne donc tout à la fois 1) le développement *d'infrastructures* et *d'outils*, tels des systèmes de stockage de données, des logiciels de base de données et de visualisation, et 2) le développement et la mise en œuvre de *méthodes* mathématiques et informatiques, s'appuyant sur ces outils, pour analyser des données et les interpréter biologiquement. Ces deux aspects de la discipline sont d'ailleurs souvent désignés sous deux noms différents en anglais, la « *bioinformatics* » couvrant le développement d'infrastructures et d'outils tandis que le « *computational biology* » désigne la mise au point de méthodes d'analyse spécifique et leur utilisation pour traiter un problème biologique particulier. En schématisant à l'extrême, la « bioinformatics » se rapproche plus d'un travail d'ingénierie en infrastructure et logiciel, pas toujours spécifique d'ailleurs au domaine d'application que sont les sciences du vivant, tandis que la « computational biology » s'apparente à une

[2] 1 pétaoctet = 1.000 téraoctets = 1.000.000 gigaoctets = 1.000.000.000 mégaoctets

discipline scientifique à part entière, consistant à modéliser et analyser des systèmes biologiques par des modèles informatiques, en empruntant d'ailleurs de nombreuses approches aux mathématiques et à la physique.

Les applications industrielles de la bio-informatique

La bio-informatique dans son ensemble est donc une activité transverse qui peut être appliquée à de nombreux secteurs des sciences de la vie et des biotechnologies confrontés à l'utilisation et l'étude du vivant. Elle joue de ce fait un rôle important et croissant dans de nombreuses industries allant de la recherche biomédicale à l'agro-alimentaire, en passant par l'énergie et l'environnement.

Les entreprises pharmaceutiques et biotechnologiques sont certainement les premières utilisatrices en volume de la bio-informatique. En effet, elles utilisent de plus en plus de technologies à haut débit, comme la protéomique ou le séquençage, pour étudier les systèmes biologiques qui les intéressent. Elles s'appuient naturellement sur les outils et les méthodes de la bio-informatique pour décoder l'information biologique cachée dans les multiples données qu'elles génèrent, et ainsi faciliter la traduction de ces données en avancées médicales. Un domaine particulier au cœur de la révolution en cours est la pharmaco-génomique, qui vise à prédire la probabilité qu'un individu réponde à un traitement en fonction de son patrimoine génétique ou de marqueurs moléculaire, ouvrant ainsi la voie à la médecine personnalisée. Cette discipline s'appuie sur des traitement informatiques et mathématiques précis, nécessitant des statistiques en grande dimension et de la fouille de données, pour identifier les combinaisons de marqueurs permettant de diagnostiquer avec précision une pathologie et de prédire l'efficacité et la toxicité d'un traitement sur un individu donné. Les enjeux sociétaux et économiques de la médecine personnalisés sont considérables, puisqu'il s'agit d'améliorer la sûreté et l'efficacité des traitements en prenant en compte les spécificités moléculaires de chaque individu.

La bio-informatique joue également un rôle important pour la identifier de nouvelles cibles thérapeutiques, correspondant à des molécules (typiquement des protéines) dont l'inhibition par un traitement permettrait de traiter la pathologie. Elle fournit d'une part des outils pour interpréter systématiquement les résultats d'études visant à caractériser les variations moléculaires entre individus (anomalies génomiques, différences au niveau de l'expression des gènes ou des marqueurs épigénétiques, etc…) et à les corréler avec l'apparition et le développement de certaines maladies. Ce travail peut permettre d'identifier des anomalies, au niveau moléculaire, responsables du développement de la pathologie, et d'en déduire des stratégies thérapeutiques comme l'inhibition d'une protéine mutée conférant une propriété particulière à une cellule cancéreuse. Une autre approche, complémentaire, pour identifier de nouvelles cibles thérapeutiques est de modéliser mathématiquement le fonctionnement d'une cellule dans un environnement donnée et, par simulation et analyse du modèle, d'en déduire des interventions thérapeutiques possibles. Ce genre de modèles, qui est au cœur de ce que l'on appelle la « biologie des systèmes », apporte d'ailleurs bien plus que l'identification de cibles candidates : il fournit également un cadre conceptuels et computationnel permettant d'intégrer des connaissances d'experts et des données mesurées sur des échantillons biologiques, ouvrant ainsi la voie à une compréhension holistique de mécanismes parfois compliqués [4]. La modélisation de voies de signalisation cellulaires ou de réseaux de régulation, décrivant au niveau moléculaire comment une cellule réagit à son

environnement et met en place des programmes d'activité particuliers, peut ainsi aider à comprendre et à prédire comment une cellule réagirait à une ou plusieurs perturbations spécifiques, permettant d'une part d'identifier les meilleures interventions possibles et d'autre part de prédire leurs effets secondaires (Figure 1).

La bio-informatique intervient également de plus en plus et de manière multiforme, avec sa cousine la « chemo-informatique », dans le processus de recherche de médicament qui est au cœur des entreprises pharmaceutiques. La recherche de nouvelles molécules inhibant ou promouvant l'activité d'une cible identifiée, et pouvant aboutir à l'élaboration d'un nouveau médicament, est en effet un processus long et coûteux qui souffre d'une chute de productivité depuis de nombreuses années notamment parce que de trop nombreuses molécules se révèlent inefficace ou trop toxiques lors de la phase finale d'essais cliniques. Qu'il s'agisse de modéliser les interactions moléculaires 3D entre différentes molécules, afin d'identifier la meilleure molécule à synthétiser pour inhiber une cible donnée ou au contraire ne pas interagir avec une autre protéine pour garantir sa spécificité (Figure 2), ou de développer des modèles *in silico* permettant de prédire la toxicité et les effets secondaires d'une molécule avant de la synthétiser et de la tester effectivement sur des patients lors d'essais cliniques, les modèles mathématiques et les outils informatiques abondent dans l'ensemble du processus de recherche de médicament des entreprises pharmaceutiques.

A côté de ses applications dans les industries de la santé, la bio-informatique joue un rôle important et croissant dans l'ensemble des industries manipulant des organismes vivants et désireuses de les étudier, de les optimiser ou de les contrôler. L'industrie agro-alimentaire s'appuie par exemple de plus en plus sur des technologies à haut débit pour disséquer et optimiser les organismes (bactéries ou levures) qu'elle utilise pour produire des aliments, dans le but d'améliorer leur goût, leur odeur, leur texture ou leur valeur nutritionnelle. Une tendance similaire est observée dans l'optimisation des propriétés des aliments même (végétaux ou animaux) par modification génétique. Dans les deux cas, l'utilisation croissante de techniques à haut débit s'accompagne naturellement d'une utilisation croissante de la bio-informatique pour analyser et exploiter ces données, ainsi que pour remplacer des expériences réelles par des simulations numériques. D'autres domaines d'application, comme les énergies renouvelables, la méta-génomique ou la biologie de synthèse, suivent évidemment la même tendance.

Le marché de la bio-informatique et ses acteurs

Compte tenu de l'importance et de la variété de ses applications, la bio-informatique s'est non seulement développée comme une discipline scientifique, mais également comme un secteur industriel à part entière et en forte croissance depuis 15 ans. Aujourd'hui on peut grossièrement segmenter le marché de la bio-informatique en trois sous-marchés principaux : 1) les logiciels d'analyse et les services associés, 2) les contenus, et 3) les infrastructures. Le marché des logiciels d'analyse, et les services associés, comprend la fourniture de solutions pour analyser et exploiter les données générées par les utilisateurs, comme par exemple les données de séquençage ou de protéomique étudiées dans la recherche pharmaceutique. Le marché des contenus recouvre de nombreuses bases de données, plus ou moins spécialisées, permettant à leurs utilisateurs d'avoir accès à des connaissances, comme par exemple les cartes des réseaux biologiques ou les informations sur les gènes et les variations génomiques

fréquentes. Enfin, le marché des infrastructures se concentre sur l'élaboration de solutions permettant à un laboratoire ou une entreprise de stocker ses données, d'en fournir l'accès aux utilisateurs, et de permettre leur analyse en terme de stockage, de puissance de calcul et de réseaux.

Le marché global total de la bio-informatique est passé d'environ 840 millions de dollars en 2002 à environ 3 milliards de dollars en 2010, suivant une croissance régulière d'environ 25% par an [5]. Plusieurs analystes considèrent que ce taux de croissance sera maintenu au moins dans les 5 prochaines années, compte tenu des investissements importants des industriels concernés dans les technologies à haut débit. En terme de volume, le marché de la bio-informatique est actuellement dominé par le secteur des contenus, suivi des logiciels et services d'analyse, suivi des infrastructures. La croissance du marché des logiciels et services d'analyse est cependant la plus forte des trois segments.

Les acteurs de ce marché sont nombreux et variés, car les coûts d'entrée sont relativement faibles et les besoins multiples. On y retrouve aussi bien des géants des technologies de l'information comme IBM, qui a créé sa division « life science » dès 2000, des grandes entreprises plus spécialisées en bio-informatique comme Accelrys (600 employés, 150 millions de dollars de chiffe d'affaire), que des myriades de petites et moyennes entreprises proposant souvent des solutions ou des services spécifiques dans des marchés très porteurs, comme SoftGenetics, DNAStar, DNAnexus ou NextBio. D'autres acteurs importants sont les grandes entreprises pharmaceutiques ou de biotechnologie elle-même, qui ont souvent investi dans des équipes de bio-informatique en propre, plus ou moins développées, et les fournisseurs de technologies pour la biologie comme Affymetrix ou Applied Biosystems, qui vendent de plus en plus des solutions bio-informatique pour récupérer, stocker et analyser les données produites par leurs technologies. Enfin, de nombreuses sociétés de biotechnologie intègrent des offres de bio-informatique en complément des produits qu'elles vendent pour la biologie, comme Kinexus (Canada) qui vend des produits et des prestations de protéomique et de bio-informatique.

D'un point de vue géographique, les Etats-Unis dominent largement le marché mondial devant l'Europe. Certains pays d'Asie sont en très forte croissance, comme l'Inde, qui bénéficie d'un très grand vivier d'informaticiens de qualité. Malgré un effort visible pour développer la bio-informatique dans le secteur académique depuis une dizaine d'année, la France reste malheureusement en retrait par rapport à ses concurrents directs sur ce secteur industriel porteur.

Les défis

La bio-informatique est devenue une discipline omniprésente dans la recherche biomédicale, et un secteur industriel en forte croissance. Cette croissance s'explique notamment par l'augmentation des volumes de données disponibles pour étudier les systèmes biologiques, et le basculement progressif de la biologie vers une science plus quantitative. Elle répond également à la nécessité d'améliorer la productivité des industries qui l'utilisent par une meilleure exploitation de ces données. De nombreux défis scientifiques et technologiques restent cependant à relever, représentant autant d'opportunités pour l'avenir.

Tout d'abord, les volumes de données générés en biologie croissent actuellement beaucoup plus vite que les capacités de stockage et la puissance de calcul des ordinateurs. Alors que ces derniers doublent environ tous les 18 mois (selon la « loi » de Moore), le coût du séquençage a été divisé par 1000 entre 2008 et 2012 (Figure 3), et les infrastructures se sont rapidement développées. Fin 2011, le plus grand centre mondial de séquençage, basé en Chine, disposait de 167 séquenceurs capables de séquencer l'équivalent de 4000 génomes humains par jour (Figure 4)! Nous sommes entré brutalement dans une ère où le coût et les capacités de production des données s'effacent devant le coût du stockage, de la transmission et surtout de l'analyse de ces données. Alors que l'on s'achemine rapidement vers la possibilité de séquencer un génome humain pour 1000$[3], de nombreux experts ont mis en avant récemment les coûts réels beaucoup plus élevés si l'on prend en compte toute la chaîne de traitement des données nécessaire à leur exploitation. C'est ainsi que Bruce Korf, ancien président de l'American College of Medical Genetics, parle de « the $1000 genome, the $1 million interpretation » [6]. Des géants de l'Internet ont commencé à s'intéresser aux problèmes de stockage et de dissémination de l'information biologique, comme Amazon qui met à disposition de la communauté scientifique les données du projet « 1000 genomes » (correspondant aux génomes séquencés de 1000 individus) sur son service de cloud[4]. Il reste cependant évident que les questions de stockage, de transmission et d'algorithmes capables de s'adapter à la croissance vertigineuse des volumes que nous observons actuellement vont constituer un fantastique défi à relever.

Par-delà les questions techniques liées à l'accroissement des volumes de données, de nombreuses questions plus scientifiques sur la manière d'analyser les données restent ouvertes. Que faire une fois que l'on aura cartographié systématiquement toutes les différences au niveau moléculaire entre des milliers d'individus ou d'échantillons biologiques ? La dernière décennie nous a montré à maintes reprises que la biologie est une science d'autant plus complexe que l'on observe la réalité à une échelle fine, et les modèles mathématiques permettant de modéliser cette complexité restent en grande partie à inventer.

Références

[1] M. R. Stratton et al. The cancer genome. *Nature* 458 :719-724, 2009.

[2] S. G. Tringe et al. Comparative metagenomics of microbial communities. *Science*, 308(5721) :554-557, 2005.

[3] B. Neumann et al. Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. *Nature* 464 :721-727, 2010.

[4] E. Barillot et al. Computational systems biology of cancer. CRC Press, 2012.

[5] RNCOS E-Service Pvt. Ltd. Bioinformatics market outlook to 2015. March 2012.

[6] K. Davies. The $1,000,000 genome interpretation. *Bio-IT World*, oct 2010.

[3] Le séquençage du premier génome humain en 2001 a coûté 2,7 milliards de dollars.
[4] http://aws.amazon.com/1000genomes

Figure 1 :

La biologie des systèmes cherche à comprendre les propriétés biologiques en appréhendant la complexité des interactions moléculaires. Cette image montre par exemple une représentation schématique d'un réseau de régulation, décrivant comment certains gènes contrôlent l'activation ou l'inhibition d'autres gènes.

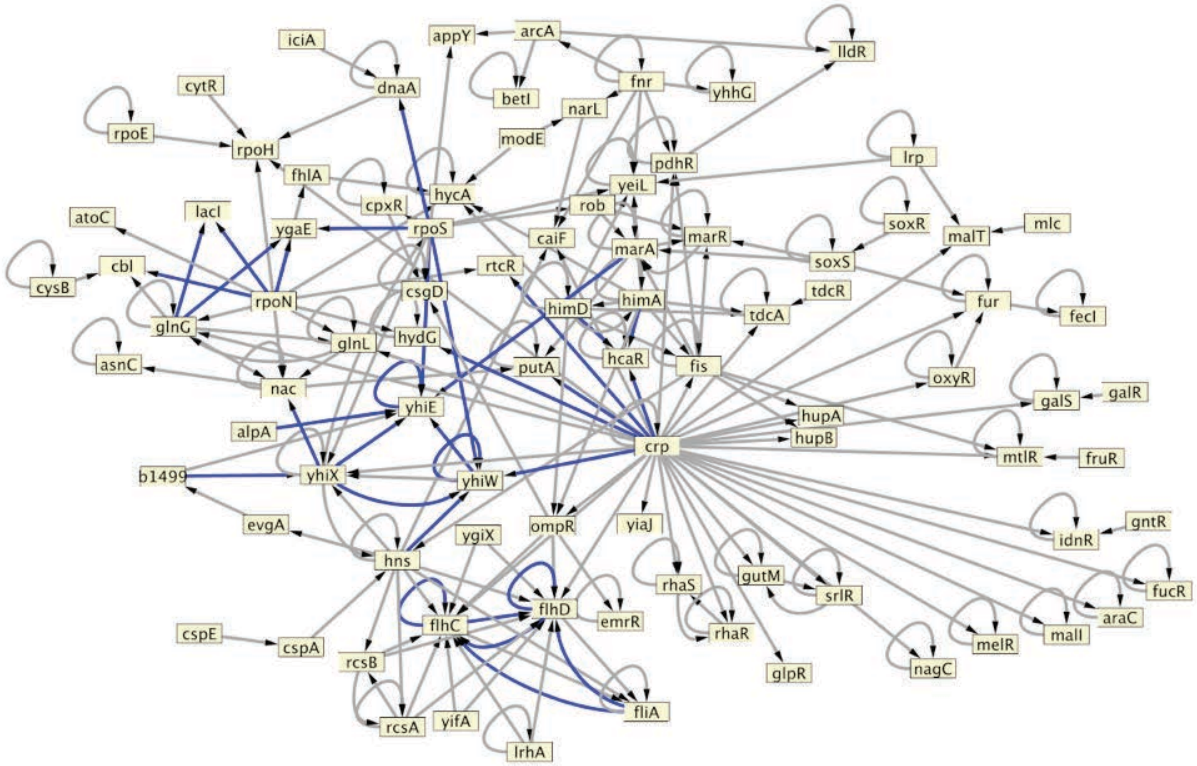

Figure 2 :

Dans l'industrie pharmaceutique, les simulations 3D permettent d'optimiser virtuellement les structures des molécules permettant d'inhiber une cible donnée, ici une protéine de la famille des tyrosine kinases qui permet à certains cancers de se développer.

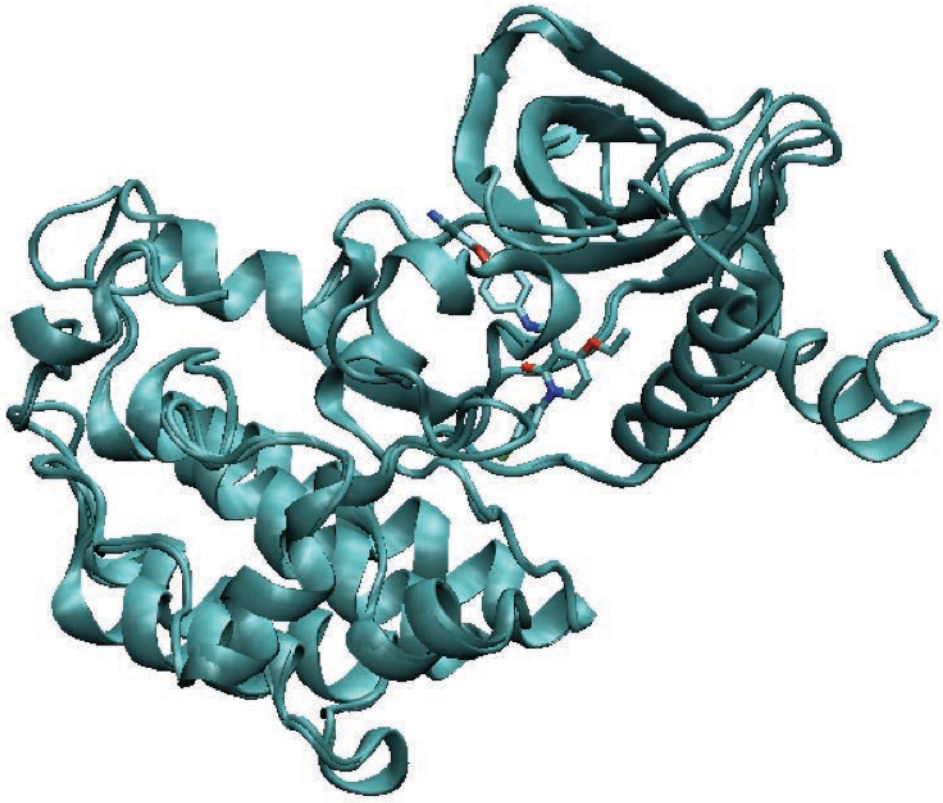

Figure 3

Le coût de séquençage d'un génome humain est passé de presque 100 millions de dollars en 2001 à moins de 10.000 dollars en 2010. La brusque décroissance des prix à partir de 2007 correspond à l'arrivée sur le marché de techniques de séquençage massivement parallèles. En comparaison, la loi de Moore décrivant l'amélioration des performances des ordinateurs fait piètre figure. (Source : KA Wetterstrand, NHGRI).

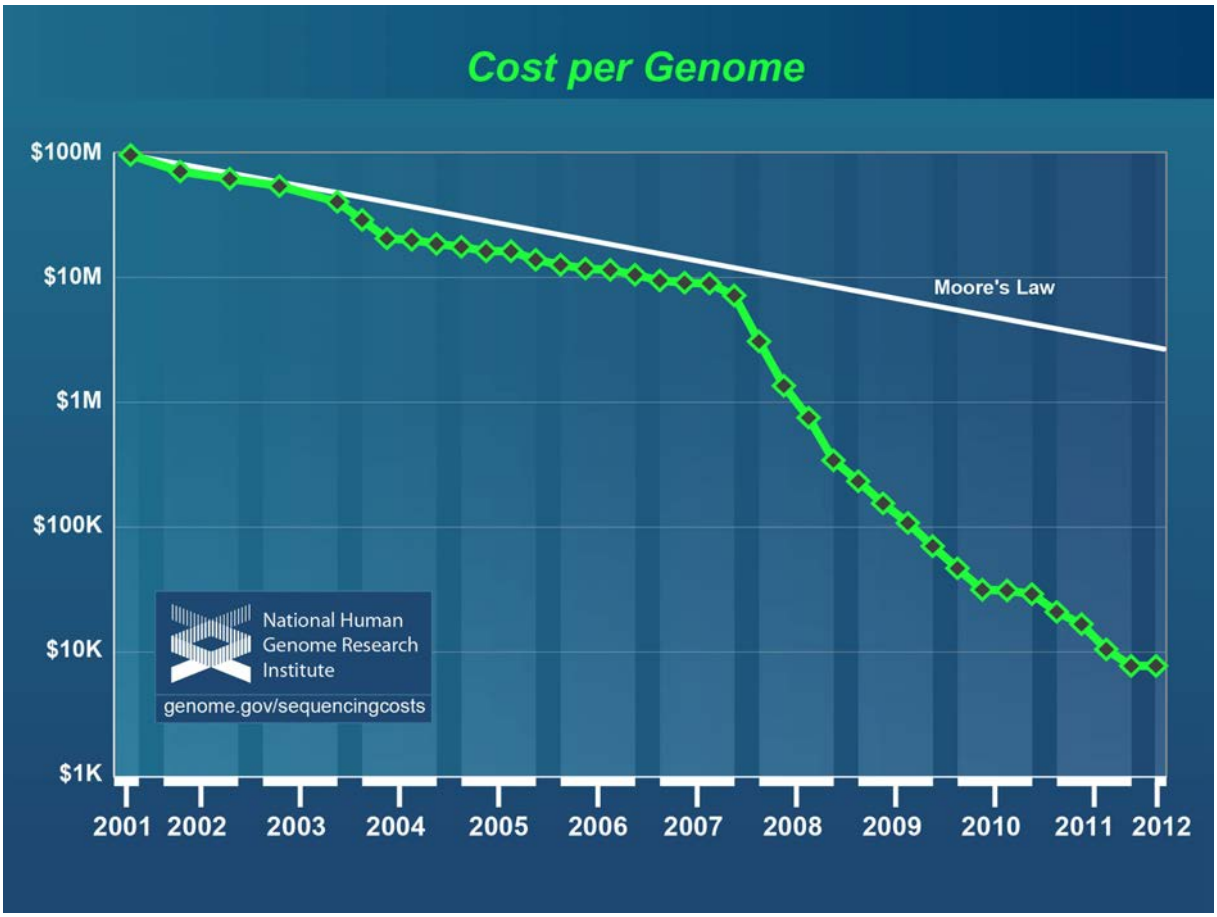

Figure 4

L'institut BGI, basé en Chine, est le plus grand centre de séquençage du monde. Il possède des centaines de séquenceurs de dernière génération, capables de générer plus de 5 téraoctets de données par jour, soit l'équivalent de 4000 génomes humains… (Source : BGI).